Environmental Data Laboratory
Professor Kirchner

Laboratory 7: Fitting functions to data, part 2

In last week's lab, you explored some of the problems that can arise if one fits an excessively complex function to data that cannot adequately constrain it. You tried to describe a noisy data set (the Venice sea level time series) using functions of differing degrees of complexity. You observed (we hope!) that more complex functions can yield a better fit, as measured by the goodness-of-fit parameter r^2, even though these more complex functions are actually worse predictors of the future behavior.

In this week's lab, you will explore the flip side of this issue, namely: how can we tell when a function we have used to fit the data is actually too simple?

Here's the hypothetical problem you'll be trying to solve (the data are actually real, though they come from a slightly different situation): Your job is monitoring concentrations of a contaminant in the Bay. Concentrations near some sewage outfalls can range up to several ppm (several thousand ppb), although more typical concentrations are in the range of 10-100 ppb. Your instrument for measuring this contaminant is an Atomic Absorption Spectrophotometer (or AA for short), which measures the absorption of a distinctive spectral line that is characteristic of the element in question. What the AA actually delivers are values of "absorbance", which describes how much of the light in a particular wavelength is attenuated by the element.

Like virtually all instruments used for environmental monitoring, the AA must be "calibrated" or "standardized" for its results to be useful. To do this, you take readings on a series of known standards, and from the absorbance readings for your known standards you construct a "calibration curve", a function that expresses how absorbance readings are related to concentrations, that is:

concentration = f(absorbance)

Where f may be a linear function, or it may be a quadratic, or it may be something more complicated still. Once the calibration function is determined from the standards, it can be used to estimate the concentrations of the "unknowns" (that is, the real-world samples) from their absorbance readings. (NB: It goes without saying that calibrated measurements are only as good as the calibration standards that they're based upon. If your calibration standards are contaminated, or otherwise inaccurate, you're utterly sunk. That's why analytical chemists are so picky about their calibration standards.)

Being extra careful, you wisely have two sets of standards, ranging from 10 ppb to 10 ppm (or 10,000 ppb). You run both sets of standards through the AA. The absorbance readings are recorded in a JMP file, "AA standards"; open this file now.

To construct a calibration curve for your standards, you first plot the concentration of the standard (on the y-axis), as a function of the absorbance reading (on the x-axis). Do this now, using "fit Y by X". A common issue with analytical instruments is so-called "linearity": whether the instrument's readings are a linear function of the concentration being measured. •1: Does it appear that your AA is linear, over the range of the standards? Do your two standards agree with one another?

Fit a straight line to the data, using the "Bivariate fit" triangle. •2: Does the line appear to fit the data well? Goodness-of-fit is often measured in terms of r^2; r^2=1 is a perfect fit and r^2=0 is a very poor fit. •3: Does the r^2 suggest that the linear fit is good? •4: Write down the linear equation, which should be in the form "concentration=a+b*absorbance, where a and b are the estimates (see the "parameter estimates" window) for the intercept and the absorbance term (absorbance is called "abs" in the data table).

Keep the calibration plot window open; you'll need it again in a minute.

Now that you have your calibration function, let's apply it to some real-world measurements. The file, "AA unknowns" contains the absorbance readings for a series of samples that you measured just after your standards. Open this file and look at the data for a moment. Note that each of the sites has two replicate samples (you've been smart: having two sets of replicate samples, like two sets of standards, helps protect you against flukes).

Now, in the "unknowns" file, construct a new column that contains your calibration function and converts the absorbance readings to concentrations. Plot the distribution of the resulting concentrations, and •5: calculate the mean, standard deviation, and standard error.

Now, why is the average of the measured concentrations *negative*? Why, indeed, are over half the measured unknowns negative? •6: If you have any ideas why this might be happening, jot them down now.

One possibility is simple random variability; sometimes the instrument reads a little high, sometimes it reads a little low. So, is that the reason? One way to test this is to see how consistent the replicate readings are. •7: Are your two sets of replicate unknowns consistent with one another? Another possibility is that there's something awry with your standards. •8: Check again: are the standards consistent with one another? Now, you've got a puzzle on your hands. If you're trying to uphold a reputation for accuracy, this is more than a little embarrasing. Your measurements seem perfectly reproducible...the only problem is that concentrations below zero are physically impossible!

If all your readings are reproducible, then the source of the problem almost has to be the way that you converted those readings into concentrations. Could there be something wrong with your calibration curve? One way to check is to calculate an estimate of the concentration for each of your known standards, using your calibration function. Do this now, and compare the calibrated concentrations to the known concentrations of the standards. •9: Do they agree? If not, in what way do they differ?

Construct a new column, called "residuals", which are the true concentrations of your standards, minus the calibrated concentrations. The residuals, then, measure the misfit (literally the "mis-fit"--here the colloquial language is right on the mark) between the calibration curve and the standards. Now, plot the residuals as a function of absorbance. •10: Do the residuals seem to be randomly scattered around, or is there a pattern to them? That is, does the calibration curve deviate from the standards in some orderly, systematic way, or does it do so randomly? In which plot do you see the mismatch between the calibration and the standards more clearly: the plot of the residuals, or your original plot for the calibration curve?

How great is the "mis-fit" between the calibration curve and the standards, compared to the concentration of the standards themselves? To see this most clearly, construct yet another column that expresses the residuals as a percentage of the concentration of each standard. •11: Given the discrepancies between the calibrated concentrations and the true concentrations, can you explain why your calibration line still appears to fit the data so well? (Hint: are the discrepancies between the calibrated concentrations and the true concentrations large, or small, compared to the range of the standards?).

Note that the typical values of your unknowns are much smaller than some of your standards. How well does your calibration curve fit the standards within the typical range of the *unknowns* (as compared to the range of the *standards*)? To explore this, expand the lower left corner of your original calibration plot. Double-click on the x-axis numbers, and select a new range for the x-axis, one that is similar to the range of the absorbances of the unknowns (in other words, crop the standards higher than this out of the plot). Then double-click on the y-axis numbers, and select a new range for the y axis so the remaining values of the standards are spread out on the plot. We suggest a range of about 0 to 0.1 on the x-axis and -100 to 350 on the y-axis. Leave this plot up, and construct another full-scale plot of the calibration curve, and look at them side-by-side. •12: Can you explain why the full-scale plot of the calibration curve looks so good, even though the calibration curve fits so poorly over the smaller range of absorbance where all your unknowns lie?

One possibility is that your AA is not quite linear enough. Let's explore what happens when you fit a more complicated curve to your data. Go to your full-scale plot of the calibration curve, and use the "fitting" option to superimpose a quadratic curve on top the linear calibration curve that's already there. •13: Is there any apparent difference between the quadratic and the linear calibration curve? Has the r^2 improved by a lot? Why or why not? Now, go to your expanded plot of the lower corner of the curve, and add a quadratic curve (JMP will fit it to all the

data, not just the points that you can see in the plot). •14: Does the quadratic calibration curve fit the low-concentration standards much better than the linear calibration curve? Write down the formula for the fitted quadratic calibration curve.

Now, as before, calculate the calibrated concentrations for the standards, using the quadratic calibration curve. There's a quicker way to do this than by typing in the formula: from the "polynomial fit" triangle just below the plot, pull down "save predicteds". Also calculate the residuals, which you can also do by pulling down "save residuals" from the "polynomial fit" triangle. Plot the residuals as a function of absorbance. •15: How does the magnitude of the residuals compare with those from the linear calibration curve? Is there any pattern to the mismatch between the calibration curve and the standards (i.e., the residuals)? Are the mismatches one-sided, or do they straddle zero? •16: By hand, sketch the residuals from the linear and quadratic calibration curves on the same axes, distinguishing them by two separate symbols.

Now, use your quadratic calibration curve to evaluate the concentration of your unknown samples. •17: Record the average concentration for your unknowns, the standard deviation, and the standard error of the mean. From the standard error of the mean, could you have anticipated how much the concentrations would differ between this calibration curve and the linear one (step 5)?

Note that the residuals display the deviation of the data from the fitted line, and they therefore hold the key to uncovering any additional information in the data that hasn't been captured by the line. For example, plot the residuals from the linear and quadratic calibration curves as functions of the absorbance. Now, look at the residuals from the linear calibration. Is there a quadratic signal there (i.e., a parabola)? To check, try fitting a quadratic to the residuals. Does it fit well? •18: If you add this quadratic to the original linear curve, do you get the quadratic calibration curve? Now try fitting the residuals from the quadratic calibration curve, using a cubic function. •19: Is there any evidence of a "cubic" signal in the residuals that should be included in the calibration curve?

Now, finally, look at the JMP data file AA standards-element #2. This file contains measurements on calibration standards for another element. •20: By fitting, examining the residuals, fitting to the residuals, and so forth, determine whether a linear, quadratic, or cubic calibration curve is most appropriate for this element. Explain your choice (note that r^2, by itself, is poor justification for choosing one formula over another).

<u>Here's what we hope you've learned from this lab, and from the session last week:</u>

1. Whether a curve fits the data "well", or whether it fits "well enough", depends on what you are using the fit for. For a calibration curve, the right question is not, "is the fit good, over the range of my standards," but instead "is the fit good, over the range that my unknowns will fall in, and is the curve reasonable between my standards in that range?" With the Venice sea level data, the right question was not, "which curve fits the data most closely?", but rather "which curve best fits the underlying trend that's hiding underneath these noisy data?"

2. You must actually <u>look</u> at your data, at scales that are relevant to the problem at hand. Statistical parameters like r^2 can sometimes be useful, but they can also be misleading. Visual checks are indispensible.

3. The value of a close fit to the data depends on how noisy the data are. The Venice sea level data contained a substantial component of noise. If you tried to fit those data too closely, you wound up fitting your curve to the random noise of the data rather than the underlying signal. In today's example, there is a very clear signal that is confounded by only the tiniest amount of random noise, so you can fit your curve much more closely to the data.

4. Residual plots are a massively useful tool for revealing the discrepancies between theory and data. A systematic pattern in the residuals indicates a systematic pattern in the data that is not accounted for by the theory. On the other hand, if the residuals show only a random pattern, then there's probably no more information left in the data that the theory hasn't already captured.